Identification methods of changes in variance in processes

M. Balcerek
Wrocław University of Science & Technology

Angers, 23 July 2016

Motivation

During cooperation with Davide Calebiro's group from Würzburg, there was a question asked: how to identify point in which process changes some parameters? After some initial analyses we found out that the most interesting parameter is the variance of the process. Specifically, in case of standard diffusion (e.i. Brownian Motion with variance $2D \cdot t$ in time t), the parameter D has an important meaning.

Analysing MSD

The most popular method of analysing SPT is anylyis of mean-squared displacement (MSD). For one-dimensional case, one can calculate it using the following formula:

$$\left\langle (x(\tau + t) - x(t))^2 \right\rangle = 2D\tau,$$

where $\langle \ldots \rangle$ is an temporal average (taken on the whole trajectory), and τ is a difference between measurements.

LLR - Log-likelihood ratio

Let's consider one dimensional diffusion and assume that we measure the position of observed particle every δ seconds. Let $\{x_i\}, \text{ for } 0 \leq i \leq N$ denote the position of the particle in for times $t_i = i\delta, i = 0, 1, \ldots, N$. Assuming a constant diffusion coefficient during the whole experiment, the increments $\Delta_i \equiv x_i - x_{i-1},\ i = 1, \ldots N$ are independent and have Gaussian distribution with mean 0 and variance $2D\delta$. Thus, the join density is equal:

$$f(\{\Delta_i\}_{i=1,\ldots,N}; D) = \prod_{i=1}^{N} \frac{1}{\sqrt{4\pi D\delta}} \exp\left[-\frac{\Delta_i^2}{4D\delta}\right]$$

Using this formula:

$$f(\{\Delta_i\}_{i=1,\ldots,N}; D) = \prod_{i=1}^{N} \frac{1}{\sqrt{4\pi D\delta}} \exp\left[-\frac{\Delta_i^2}{4D\delta}\right]$$

we can calculate the MLE $\hat{D}$ of parameter D:

$$\hat{D} = \frac{1}{2N\delta} \sum_{i=1}^{N} \Delta_i^2$$

LLR - Log-likelihood ratio

We are testing the hypothesis:

$$\begin{aligned} H_0: &\quad D(t_1) = D(t_2) = \ldots = D(t_N) = D_0, \\ H_1: &\quad D(t_1) = D(t_2) = \ldots = D(t_i) = D_1 \\ &\quad \neq D_2 = D(t_{i+1}) = \ldots D(t_N). \end{aligned}$$

To test H_1 against H_0 we will calculate the logarithm of appropriate likelihood functions.

LLR - Log-likelihood ratio

$$2\mathrm{llr}(k) = 2\ln\left[\frac{f(\{\Delta_{i=1,\dots,k}|\hat{D}_1\})f(\{\Delta_{i=k+1,\dots,N}|\hat{D}_2\})}{f(\{\Delta_{i=1,\dots,N}|\hat{D}_0\})}\right] =$$

$$= N\ln[\hat{D}_0] - k\ln[\hat{D}_1] - (N-k)\ln[\hat{D}_2],$$

where

$$\hat{D}_0 = \frac{1}{2N\delta}\sum_{i=1}^{N}\Delta_i^2,$$

$$\hat{D}_1 = \frac{1}{2k\delta}\sum_{i=1}^{k}\Delta_i^2,$$

$$\hat{D}_2 = \frac{1}{2(N-k)\delta}\sum_{i=k+1}^{N}\Delta_i^2,$$

are most likelihood estimators of diffusion coefficient in the appropriate time limits.

LLR - Log-likelihood ratio

Index k that maximizes the value $\mathrm{llr}(k)$ is the most likely index of a change point in diffusion coefficient.
Hence, we will use the test statistic:

$$Z_N = \max_{1 \leq k \leq N} \{2\mathrm{llr}(k)\}.$$

Let's denote the critical value of the test (on significance level $1-\alpha$) as $C_{1-\alpha}$.

LLR - Log-likelihood ratio

$$\begin{cases} Z_N^{1/2} \geq C_{1-\alpha}, & \text{then we consider that there was a change in D} \\ & \text{in point } \hat{k} = \text{argmax}_{1\leq k\leq N}\{2\text{llr}(k)\}, \\ Z_N^{1/2} < C_{1-\alpha}, & \text{then we consider there was not a change in D.} \end{cases}$$

The values $C_{1-\alpha}$ can be calculated numerically.

MARS - Multivariate Adaptive Regression Splines

MARS method was developed by Jerome Friedmana (Stanford University) in the 90's. The main point of the method is to fit a function of the following construction to the data.

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x).$$

Model is a weighted sum of base functions $B_i(x)$. Each of those base functions $B_i(x)$ has one of the following forms:

- $B_i(x) = 1$, e.i. $B_i(x)$ is a constant;
- $B_i(x) = \max(0, x - \alpha_i)$ or $B_i(x) = \max(0, \alpha_i - x)$;
- $B_i(x)$ is a product of the previous types functions.

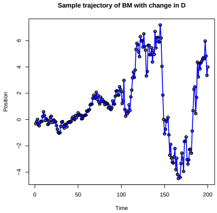
MARS - Multivariate Adaptive Regression Splines

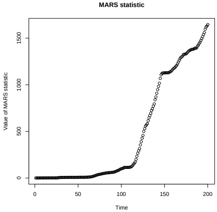
Given the observed increments $(x_1, \ldots, x_N)$ let's denote by m their median. Using the statistics:

$$C_j = \sum_{i=1}^{j} |x_i - m|$$

and the J.Friedman's method, we find parameters of function $\hat{f}$. Found parameters α_i point to possible change points in variance. After that, we still need to test, if the points are significant. One can use, e.g. Ansari-Bradley test (it checks if both samples come from the same distribution, against the alternative that they come from distributions with the same median and shape but different variance).

MARS - Multivariate Adaptive Regression Splines





Regime Variance

In the Regime Variance method we calculate the following statistic:

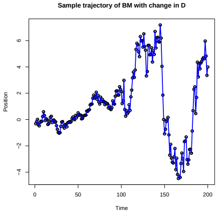
$$C_j = \sum_{i=1}^{j} x_i^2, \qquad j = 1, \ldots, N,$$

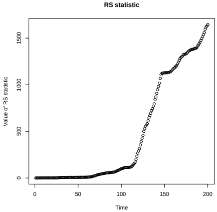
where x_i are the observed increments.
Precisely, we check the following hypothesis:

$$H_0 : \quad \text{Quantiles of order } \alpha/2 \text{ and } 1-\alpha/2 \text{ of } (x_j)_j \text{ do not change in time;}$$

$$H_1 : \quad \text{There exist at least one time point, where the sample quantiles of } \{x_1, \ldots x_k\} \text{ and } \{x_{k+1}, \ldots, x_N\} \text{ are different.}$$

Regime Variance





We need two steps to find a change point. First, we need to separate the statistics C into two sets:

$$\{C_1, C_2, \ldots, C_k\} \text{ i } \{C_{k+1}, C_{k+2} \ldots, C_N\},$$

in which, using linear regression, we will fit linear functions:

$$\begin{aligned} y_k^{(1)}(j) = & \quad a_1(k)j + b_1(k), \qquad j = 1, \ldots, k, \\ y_k^{(2)}(j) = & \; a_2(k)j + b_2(k), \qquad j = k+1, \ldots, N. \end{aligned}$$

Then, we will look for such k, that will minimize the following sum:

$$\sum_{j=1}^{k}(C_j - y_k^{(1)}(j))^2 + \sum_{j=k+1}^{N}(C_j - y_k^{(2)}(j))^2$$

Regime Variance

To check if the found point $\hat{k}$ is significant we can calculate the p-value:

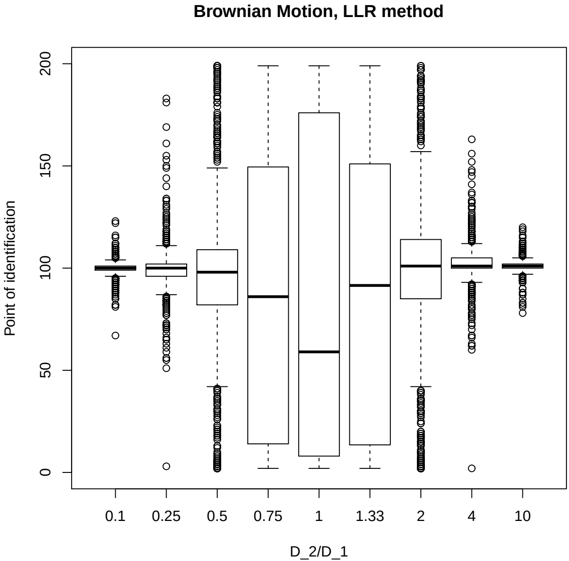
$$\text{p-value} = \mathbb{P}(Z < B),$$

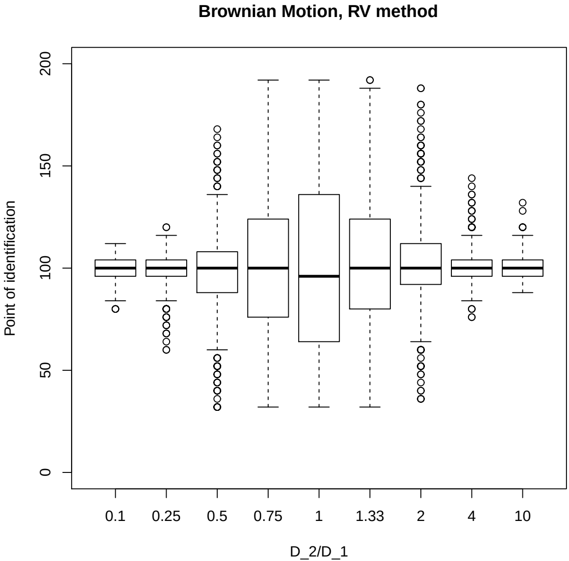
where Z has binomial distribution with parameters $(n - \hat{k}, 1 - \alpha)$, and $B = \#\{i : q_{\alpha/2} \leq x_i \leq q_{1-\alpha/2}\}, \text{ while } q_\alpha$ denotes the sample quantile of order α.

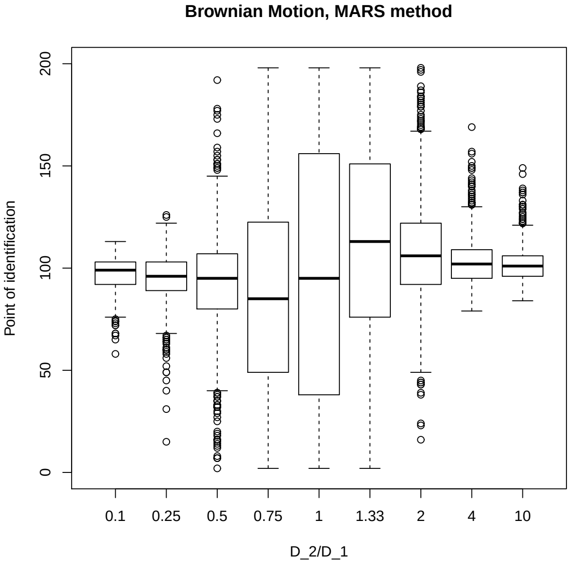
Testing methods - simulation

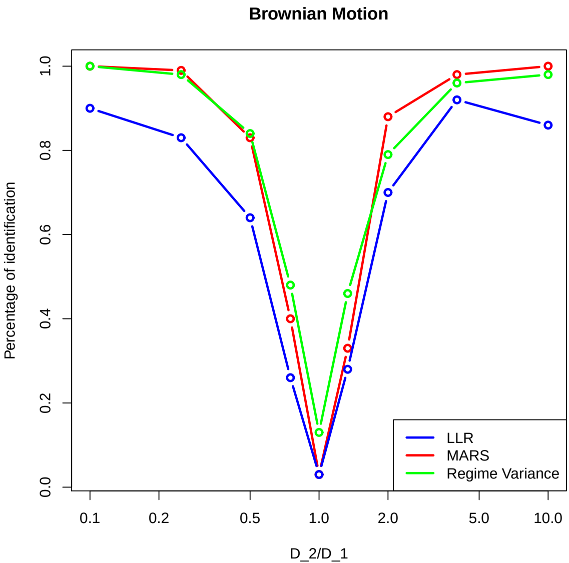
All methods were tested using Monte Carlo simulations.

1. A sample: $(x_1, x_2, \ldots, x_N)$, where $x_1, \ldots, x_k$ are independent realizations of normal r.v. $\mathcal{N}(0, D_1)$ and $x_{k+1}, \ldots, x_N$ are independent realizations of normal r.v. $\mathcal{N}(0, D_2)$. The ratio of variances between both parts is equal to: $\frac{D_2}{D_1}$. The assumed parameters were: $N = 200, k = 100, D_1 = 1$.


Brownian Motion, LLR method
Point of identification
0
50
100
150
200
0.1
0.25
0.5
0.75
1
1.33
2
4
10
D_2/D_1


Brownian Motion, RV method
Point of identification
0
50
100
150
200
0.1
0.25
0.5
0.75
1
1.33
2
4
10
D_2/D_1


Brownian Motion, MARS method
Point of identification
0
50
100
150
200
0.1
0.25
0.5
0.75
1
1.33
2
4
10
D_2/D_1


Brownian Motion
Percentage of identification
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.5
1.0
2.0
5.0
10.0
D_2/D_1
LLR
MARS
Regime Variance

Conclusion

- All presented methods work quite well in case of Gaussian i.i.d. r.v.;
- A big positive - MARS method is available in an R package;
- LLR and RV methods are easily implemented on computer;
- LLR and MARS methods work very fast, RV is slower.

What's next?

- Computing LLR method for fractional Brownian motion or maybe some stable distribution (scale parameter);
- Analysing how the performance is influenced by length of our time series.

Merci beaucoup de votre attention!

Sources I

- D. Montiel, H. Cang & H. Yang
Quantitative Characterization of Changes in Dynamical Behavior in Single-Particle-Tracking Studies.
Journal of Physical Chemistry B, 110, 19762–19770 (2006).

- J. Gajda, G. Sikora & A. Wyłomańska
Regime Variance Testing - a Quantile Approach.
Acta Physica Polonica B, 5, Vol. 44 (2013).